Published in final edited form as:
J Comput Biol. 2008 September ; 15(7): 857–866. doi:10.1089/cmb.2007.0148.

Estimating Genome-wide Copy Number using Allele Specific Mixture Models

Wenyi Wang,
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

Benilton Carvalho,
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

Nathaniel D. Miller,
Department of Neurology, Kennedy Krieger Institute, 707 North Broadway, Baltimore, MD 21205, USA

Jonathan Pevsner,
Department of Neurology, Kennedy Krieger Institute, 707 North Broadway, Baltimore, MD 21205, USA

Aravinda Chakravarti, and
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore MD 21205, USA

Rafael A. Irizarry
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

Abstract

Genomic changes such as copy number alterations are one of the major underlying causes of human phenotypic variation among normal and disease subjects. Array comparative genomic hybridization (CGH) technology was developed to detect copy number changes in a high-throughput fashion. However, this technology provides only a >30 kb resolution which limits the ability to detect copy number alterations spanning small regions. Higher resolution technologies such as single nucleotide polymorphism (SNP) microarrays allow detection of copy number alterations at least as small as several thousand base pairs. Unfortunately, strong probe effects and variation introduced by sample preparation procedures have made single-point copy number estimates too imprecise to be useful. Various groups have proposed statistical procedures that pool data from neighboring locations to successfully improve precision. However, these procedure need to average across relatively large regions to work effectively thus greatly reducing resolution. Recently, regression-type models that account for probe-effects have been proposed and appear to improve accuracy as well as precision. In this paper, we propose a mixture model solution, specifically designed for single-point estimation, that provides various advantages over the existing methodology. We use a 314 sample database, to motivate and fit models for the conditional distribution of the observed intensities given allele specific copy number. We can then compute posterior probabilities that provide a useful prediction rule as well as a confidence measure for each call. Software to implement this procedure will be available in the Bioconductor oligo package (http://www.bioconductor.org).

Correspondence to: Rafael A. Irizarry.

1 Introduction

High-resolution measurements for chromosomal copy number estimates can be obtained using SNP microarray platforms such as those developed by Illumina and Affymetrix or array CGH platforms from Nimblegen and Agilent [Peiffer et al. 2006, Gribble et al. 2006, Sharp et al. 2006]. Statistical methodology has been developed for Affymetrix SNP arrays to provide copy number estimation algorithms [Zhao et al. 2004, Bignell et al. 2004, Huang et al. 2004, Nannya et al. 2005, Ishikawa et al. 2005, Komura et al. 2006, Huang et al. 2006, Laframboise et al. 2006]. An advantage of the SNP chip technology is that we can obtain genotype calls which permits allele specific copy number estimation. We can then use these to predict parent specific copy number that is useful in detecting uniparental disomy [Laframboise et al. 2006].

The genotyping platform provided by Affymetrix interrogates hundreds of thousands of human single nucleotide polymorphisms (SNPs) on a microarray. DNA is obtained and fragmented at known locations so that the SNPs are far from the ends of these fragments; the fragmented DNA is amplified with a polymerase chain reaction (PCR); and the sample is labeled and hybridized to an array containing probes designed to interrogate the resulting fragments. We refer to the measurements obtained from these probes as *feature intensities*. There are currently three products available from Affymetrix: an array covering approximately 10,000 SNPs (GeneChip Human Mapping 10K), a pair of arrays covering approximately 100,000 SNPs (GeneChip Human Mapping 50K Xba and Hind Array), and a pair of arrays covering approximately 500,000 SNPs (GeneChip Human Mapping 250K Nsp Array and Sty Array). These are referred to as the 10K, 100K, and 500K chips respectively. Affymetrix has recently launched SNP array 6.0 that contains over 900K SNPs, as well as over 900K non-polymorphic probes for the detection of copy number variation.

To motivate the model and estimation procedures described here we need to understand the basics of the feature-level data. We provide the essential details here and refer readers to Kennedy et al. [2003] for a complete description. Each SNP on the array is represented by a collection of probe quartets. As with Affymetrix expression arrays, the probes are defined by 25-mer oligonucleotide molecules referred to as perfect match (*PM*) probes. There are also mismatch probes *MM* which we completely ignore because the manufacture has plans of no longer using them. *PM* probes for SNP arrays differ from expression arrays in three important ways. First, two alleles are interrogated (for most SNPs only two alleles are observed in nature). These are denoted by A and B and divide the probes into two groups of equal size. For each *PM* probe representing the *A* allele there is an allele *B* that differs by just one base pair (the SNP). Second, features are included to represent the sense and antisense strands. This difference divides the probes into two groups that are not necessarily of the same size. Finally, for each allele/strand combination, various features are added by shifting the position of the SNP within the probe. The position shift ranges only from -4 to 4 bases, therefore within each strands the probes are relatively similar.

Most copy number algorithms can be divided into three main steps which we refer to as 1) the *preprocessing* step, 2) the copy number *estimation* step, and 3) the *smoothing* across the chromosome step. In the preprocessing step we summarize feature intensities into two quantities, representative of allele *A* and *B*. We refer to this step as *preprocessing*. In this paper, we use the following notation to denote the preprocessed data: $\theta_{A,i,j}$ and $\theta_{B,i,j}$ are the logarithms (base 2) of quantities proportional to the amount of DNA in target sample *j* associated with alleles *A* and *B* for SNP *i*. In the estimation step, we use these θs to estimate the true copy number, which we denote with $C_{i,j}$. The allele specific copy number are denoted with $C_{A,i,j}$ and $C_{B,i,j}$. Notice that the total copy number is the sum of the allele specific copy numbers, i.e. $C_{i,j} = C_{A,i,j}+C_{B,i,j}$. As we demonstrate here (see Figure 1), estimates of θ_A and θ_B are, in general, not precise enough to provide useful copy number calls. Therefore, most copy number

estimation algorithms include the smoothing step in which estimates from neighboring regions are averaged to improve the signal to noise ratio. These techniques range from simple method such as running median to more complicated ones such as hidden Markov models (HMM). In Section 3 we review some of the existing methods and motivate our mixture model approach. In Section 4 we describe the mixture model approach. In Sections 5 we present results and discussion respectively. Throughout this paper we use data obtained from collaborators and public repositories which we briefly describe in Section 2.

2 Control Data

In Section 4 we describe a model that is trained using a reference set of 314 normal samples hybridized to Affymetrix's 100K array. We screened out samples not achieving the quality standard described by Carvalho et al. [2006]. Our reference set consists of 86 Hapmap samples, 124 samples from the Coriell Repositories (42 African American, 20 Asians, 40 Caucasians and 22 samples from the polymorphisms discovery panel) [Collins et al. 1998], and 104 from Chakravarti's lab. The test data was sampled from 20 *trisomy 21* samples from Pevsner's lab.

3 Previous Work and Motivation

The first algorithms we describe do not provide allele specific results. We therefore define the total copy number quantity $S_{i,j} = \log_2(2^{\theta_{A,i,j}} + 2^{\theta_{B,i,j}})$. Ideally the $S_{i,j}$ is proportional to the true log-scale copy number. Figure 1A shows data for a male sample with Down syndrome. The $S_{i,j}$'s are highly noisy and differences between chromosome 21 and X are hard to detect unless we smooth along the chromosome (in Figure 1 we show the results of running median). The lessons learned from expression arrays help understand this problem. Various authors have proposed an additive background/multiplicative model for gene expression microarray [Rocke and Durbin 2001, Huber et al. 2002, Wu et al. 2004]. Furthermore, for Affymetrix arrays, various authors [Irizarry et al. 2003, Li and Wong 2001] have clearly shown the existence of a strong multiplicative probe-specific effect. Probe-specific background noise, attributed to non-specific binding, have also been described [Wu et al. 2004]. Others [Rabbee and Speed 2006, Laframboise et al. 2006, Carvalho et al. 2006] demonstrate that similar sized effects are seen with SNP chips. These effects are strong enough to be clearly seen even after averaging the various feature intensities associated with each SNP. Extending these findings to the copy number case results in the following model:

$$\theta_{a,i,j} = \log(\beta_{a,i}\delta_{a,i,j} + \phi_{a,i}c_a\varepsilon_{a,i,j}) \quad (1)$$

with $a = A,B$ denoting allele, i identifies the SNP, j identifies the sample, β represents a SNP-specific background level, δ represents background variability, φ represents a SNP-specific probe-effect, and ε is multiplicative measurement error that usually follows a log normal distribution with mean 1. Assuming this model holds, relatively simple calculations demonstrate that large values of $\beta_{a,i}$ result in attenuation of real differences and that large variability of $\varphi_{a,i}$ across SNPs explains the large variance seen in Figure 1A.

Affymetrix's Copy Number Analysis Tool (CNAT) [Bignell et al. 2004, Huang et al. 2004] deals with the probe-effect using a simple yet effective technique. CNAT does not provide allele specific results and concentrates on estimating the overall copy number. For the preprocessing step, all feature intensities related to the SNP are therefore averaged to form $S_{i,j}$. Using dozens of control subjects, CNAT defines a SNP specific average $\bar{S}_{ig}$, standard deviation $\hat{\sigma}_{ig}$, for each genotype g=AA,AB,BB. Values $S_{i,j'}$, from any new sample j' that are called genotype g are standardized in the usual manner: $(S_{i,j'} - \bar{S}_{ig})/\hat{\sigma}_{ig}$. A predefined regression equation is then used to transform these standardized values to the copy number scale. The standardized S values are used to obtain p-values from the null hypothesis that $S = 0$ ($C = 2$). Figure 1B shows de-meaned (observed minus mean) values for the same data shown in Figure

1A. The improvement is clear and it is due to the fact that φ is partially removed from the de-meaned values. However, notice that the signal to noise ratio still appears to be small: the separation between chromosomes with known differences is far from perfect. To avoid false positives, the third step in CNAT involves looking for strings of consecutive p-values that are smaller than some predefined cut-off. Other smoothing approaches have been used. For example, Zhao et al. [2004] proposed the use of Hidden Markov models (HMM) to define the procedure implemented by dChip.

Other authors have noted that further improvements can be obtained by reducing the variance at the preprocessing step. For example, several groups [Nannya et al. 2005, Ishikawa et al. 2005, Komura et al. 2006] have used probe-sequence information, mainly GC content, and fragment length to predict and remove some of the probe-effect related variability. However, even after accounting for such factors the signal to noise ratio remains too low to make single-point copy number calls, thus these authors propose their own versions of the smoothing step.

Huang et al. [2006] noted that CNAT's mean removal approach does not fully remove the probe effect because it does not properly deal with the additive background effect β. They propose the Copy Number Analysis with Regression And Tree (CARAT) algorithm which uses a non-linear regression model, based on model (1), to account for the probe-specific effects. To estimate model parameters they use a control dataset composed of dozens of arrays. First, genotype calls are obtained and treated as known. This permits estimation of allele specific parameter estimates. For example, for allele *A* we have known values of $c_A = 0$, $c_B = 2$ (*BB* genotype), $c_A = 1$, $c_B = 1$ (*AB* genotype), and $c_A = 2$, $c_B = 0$ (*AA* genotype) and thus we can estimate $\beta_{a,i}$ and $\varphi_{a,i}$ with, for example, least squares, for each SNP *i* and each allele *a* = *A*,*B*. For a new sample, we can predict c_A, c_B using the fitted parameters and equation (1). Calls can then be based on cut-offs for the prediction of $\hat{c}_A + \hat{c}_B$. Huang et al. [2006] suggest using [0, 1.5), [1.5, 2.5], (2.5,∞) for total copy number < 2,= 2,> 2 respectively. Figure 2A shows the data used in the regression for allele A from a randomly chosen SNP. The figure demonstrates that the model works reasonably well but that the signal to noise ratio is not large enough to provide perfect accuracy (the boxplots overlap). CARAT utilized a regression tree approach in the smoothing step.

The Probe-level allele-specific quantitation (PLASQ) [Laframboise et al. 2006] procedure is similar to CARAT. Two major difference is that PLASQ fits (1) to the feature-level data and that it does not rely on external genotype calls. Although, PLASQ provides a superior model-based framework than any other approach, it is computationally challenging to implement. This is because a non-linear estimation procedure is performed at the feature-level for every SNP. Furthermore, it is difficult to adapt it to be robust to outliers and to take probe-sequence and fragment size into account.

Model based approaches such as CARAT and PLASQ provide a great advantage over previous ones: reliable confidence intervals can be computed for single-point copy number estimates. Huang et al. [2006] point out that their uncertainty assessment permits one to call a relatively large group of SNPs and keep the false positive rate relatively low. We now briefly describe a simple adaptation of these methods that provides further improvements.

Notice that all of the above described algorithms use regression-type approaches to give a continuous prediction of copy number. The current approaches rely on three assumptions that we believe are not exactly true. The first is that the linear relationship predicted by model (1). Figure 2A and 2B show that there are small but significant deviations from these models. Other SNPs (not shown) show slightly larger deviations. The second is that θ_A and θ_B are independent. This assumption is clearly not true as demonstrated by Figure 2C and Figure 3. The third assumption is that the variance of the measurement error term does not depend on allele-specific

copy number values. Figure 3 also shows this is not the case. In general, we are making convenience assumptions regarding the conditional probabilities of $(\theta_A, \theta_B)'$ given allele-specific copy number that hurt bottom-line results.

Theoretically, one can show that the best predictor of discrete classes given continuous covariates is Bayes classifier. Bayes classifier is a function of the conditional distribution of the predictors given the classes which we can not always estimate effectively. In the next section we describe how we can use the large amount of public data and equation (1) to obtain useful estimates of these conditional distributions and therefore improved copy number calls.

4 Allele-specific Mixture Model

Figure 2C shows a scatterplot for θ_A vs. θ_B across many individuals. Notice that the three genotypes are clearly seen and that the data for each cluster appears to be bivariate normal. This result can be motivated by model (1). First we assume that data summarized by SNP-RMA can be modeled similarly to probe-level data. As long as we keep sense and anti-sense probes separate this assumption should hold as the probes have very similar sequence. Now to see how model (1) is in agreement with Figure 2C, notice that for $C = 0$, then (1) reduces to $\theta = \log(\beta) + \log(\delta)$ which follows a normal distribution. If $\varphi \ll \beta$ then the following approximation $\theta \approx \log(\varphi) + \log(\varepsilon)$ suggest θ is normally distributed. Notice as well that the cluster related to the *AB* genotype appears to show correlation. This implies that ε_A and ε_B are correlated. This is in agreement with the fact that PCR should have a similar effect on DNA fragments related to the different alleles as they have the same length and almost identical sequences. These results motivated the use of a normal mixture model defined in the following way:

$$[\theta_{i,j}|\mathbf{C}_{i,j}=\mathbf{c}]=\begin{pmatrix}\gamma_{A,c_A,i}\\ \gamma_{B,c_B,i}\end{pmatrix}+\begin{pmatrix}\epsilon_{A,c_A,i,j}\\ \epsilon_{A,c_B,i,j}\end{pmatrix} \quad (2)$$

with $\theta_{i,j} = (\theta_A, \theta_B)'$, $\mathbf{C}_{i,j} = (C_{A,i,j}, C_{B,i,j})'$ representing the un-observed true copy number of alleles *A* and *B* for SNP *i* on sample *j*, $\mathbf{c} = (c_A, c_B)'$ are the possible values $\mathbf{C}_{i,j}$ can take, $(\gamma_{A,c_A,i}, \gamma_{B,c_B,i})'$ accounts for the shifts in location caused by the probe-effect, and $(\varepsilon_{A,c_A,i,j}, \varepsilon_{B,c_B,i,j})'$ is a bivariate normal error with mean 0 and copy-number-specific covariance matrix $\Sigma_{c,i}$ which is defined by the allele-copy-number-specific standard deviations $\sigma_{c_A,i}$, $\sigma_{c_B,i}$ and copy-number-pair-specific correlation $\rho_{c_A,c_B,i}$.

For this model to be useful we need reliable parameter estimates. We also need a computationally practical solution as the model is fit for each SNP (100K,500K). To do this we took advantage of the the large database of normal individuals described in Section 2. These were genotyping using the CRLMM algorithm [Carvalho et al. 2006], which has an error rate well below 1%. We therefore assume that, for these individuals, C_A and C_B are known. This make the estimaton of the paramaters in (2) straightforward: because we know **C** for all these samples we can estimate the γs by simply using:

$$\widehat{\gamma}_{A,c_A,i}=N_{c_A,i}^{-1}\sum_{\{j:\,C_{A,i,j}=c_A\}}\theta_{A,i,j},\ c_A=0,1,2 \quad (3)$$

with $N_{c_A,i}$ the number of samples with genotypes implying $C_{A,i,j} = c_A$. The covariance matrix $\Sigma_{\mathbf{c},i,s}$ is computed in a similar way, namely using the sample covariance matrix of $\theta_{i,j}$ for samples *j* implying $\mathbf{C}_{i,j} = \mathbf{c}$. [1] Because we assume the εs are normal, these sets of parameter estimates define the conditional distributions for $\mathbf{C} = (2, 0)$, (1, 1) and (0, 2). Next we assume that the behavior of the **θ**s for $C_A = c_A$ is similar for all values of C_B and vice-versa (no cross-

[1] In this extended abstract we actually use robust (to outliers) versions of these sample means and covariances

hybridization). We then infer the conditional means for **C** = (0, 0), (0, 1), (1, 0). For example the conditional mean for SNP *i* when **C** = (0, 1) will be ($\gamma_{A,0,i}$, $\gamma_{B,1,i}$). The covariance matrix is inferred in a similar way (described below).

The above procedure permits us to predict the distributions of (θ_A, θ_B) for cases with total copy number 0,1, or 2. Model (1) becomes particularly useful when trying to predict these distributions for **C** = (3, 0),(2, 1),(1, 2), (0, 3),(4, 0),(3, 1),…. We do this by first using the estimates of $\gamma_{A,0,i}$, $\gamma_{A,1,i}$, $\gamma_{A,2,i}$ as outcomes in model (1) for values of C_A= 0, 1, 2 respectively, fit (1), and obtain estimates of $\beta_{A,i}$ and $\varphi_{A,i}$, which permit us to predict $\gamma_{A,3,i}$, for C_A = 3.

We now describe how we infer $\Sigma_{\mathbf{c},i}$ for cases other than **C** = (2, 0), (1, 1),(0, 2) using the estimates we already have. For the A and B variance components (the diagonal entry) of the covariance matrix, we simply assume they depend only on c_A and c_B respectively. For $c_A > 2$ and $c_B > 2$ we assume the same variance as $c_A = 2$ and $c_B = 2$ respectively. We therefore use the estimates of the six parameters: $\sigma_{A,c_A,i}$,c_A = 0, 1, 2, $\sigma_{B,c_B,i}$,c_B = 0, 1, 2 and do not need to predict any new values. The correlation component is a bit more difficult. We assume that the correlation coefficient when $C_A > 0$ and $C_B > 0$ is the same as **C** = (1, 1). The rationale for this is that correlations are due to PCR effects being different from sample to sample. Thus if both allele fragments are present, the resulting quantities will be similar regardless of the starting quantities. When one of the two alleles is not present (PCR no longer makes it grow) we assume that the correlation for case where $C_A > 0$ but $C_B = 0$ is the same as **C** = (2, 0) and $C_A = 0$ but $C_B > 0$ the same as **C** = (0, 2). For **C** = (0, 0) we simply assume independence. With this assumption in place we can produce conditional expectations for any value of **C** given the observed **θ**s, described as follows.

With the model parameter estimates in place we are able to provide posterior probabilities for allele specific copy number. Furthermore, we can compute these posterior probabilities for total copy number:

$$[C_{A,i,j}+C_{B,i,j}=c|\theta_{i,j}]$$

$$\propto \sum_{\{\mathbf{c}:\, c_A+c_B=c\}} [\theta_{i,j}|\mathbf{C}_{i,j}=\mathbf{c}] \times [\mathbf{C}_{i,j}=\mathbf{c}]$$

where $[\boldsymbol{\theta}_{i,j}|\mathbf{C}_{i,j} = \mathbf{c}]$ is the bivariate normal distribution defined by model (3). The marginal probability of the **C** pair can be pre-specified and used to control specificity and sensitivity for any copy number value. We can obtain meaningful values by decomposing the probability into: $\Pr(C_{A,i,j} = c_A, C_{B,i,j} = c_B) = \Pr(C_{A,i,j} = c_A, C_{B,i,j} = c_B | C_{A,i,j} + C_{B,i,j} = c) \Pr(C_{A,i,j} + C_{B,i,j} = c)$. The first component relates to the proportion of each genotype in the population and can be computed using the Hardy-Weinberg Equilibrium for diploids ($C_{A,i,j} + C_{B,i,j} = 2$). The second component relates to the probability of each alteration c = 0, 1, 3, 4,… which is unknown. We recommend the user define these probabilities to control specificity and sensitivity. For the examples shown in this extended abstract we assigned equal probabilities to $C_{A,i,j} + C_{B,i,j}$ = 0, 1, …, 6.[2] Once we have calculated the probabilities above we can provide estimates of copy number by, for example, computing the expected value of $C = C_A + C_B$.

A summary of the algorithm:

1. For each array, we obtain the pre-processed probe-level log intensities from snpRMA, the pre-processing algorithm used by CRLMM. These resulting measurements are $\theta_{A,+}$,$\theta_{A,-}$,$\theta_{B,+}$,$\theta_{B,-}$ for each SNP.

[2]Remember that we perform the above calculation separately for the sense and antisense values. A final estimate of the posterior probability simply average these two values.

2. We estimate the conditional probability of these measurements, given allele specific copy number. We assume a bivariate normal for the *A* and *B* alleles at each copy number pair. This reduces the number of parameters greatly and we can estimate them precisely using a large training set. We use genotype calls to treat the allele specific copy number as known. We do this independently for sense (+) and antisense (−). More specifically:
3. We assume the prior probability for the joint distribution of C_A and C_B is a uniform distribution.
4. For a new dataset, we use the above estimates to calculate the posterior probability for C_A and C_B being 0, 1, 2, …, K (K is the maximum copy number permitted). We average the sense and antisense results.
5. Finally, we compute the posterior probability of $C_A + C_B$ being 0, 1, 2, …, K and the posterior mean of $C_A + C_B$.

5 Results

We now describe some of the applications of the hierarchical model described above. In general we refer to our procedure as the Copy Number-Robust Linear Model and Mixture Model (CN-RLMM) procedure. The robustness is attained by using medians and robust variance estimates in place of means and variances.

Figure 3 gives the SNP-specific bivariate normal distribution of **θ** for **C** = (0, 0), (0, 1), (1, 0), (0, 2), (1, 1), (2, 0), (0, 3), (1, 2), (2, 1), (3, 0), depicted in ellipses (95% confidence regions). These are estimated from the control data as described in Section 2. Figure 3A and 3B give the sense and antisense-specific distributions for a SNP on chromosome X. We observe the extrapolated distribution of **θ**s given **C** = (0, 1), (1, 0) coincide with the observed **θ**s from 45 male samples that were not used in training. Similarly, figure 3C and 3D give the sense and antisense-specific distributions for a SNP on chromosome 21 and we observe that our extrapolated distributions coincide with the observed **θ**s from 20 *trisomy 21* samples. This demonstrates that our assumptions seem to provide reasonable estimates of the conditional distribution of copy number the cases predicted with mode (1) (**C** = (0, 0), (0, 1), (1, 0), (0, 3), (1, 2), (2, 1), (3, 0)).

In Figure 4A and 4B we demonstrate how our results have much better precision than CNAT and values with and without probe- sequence and fragment length corrections. We achieve this precision without any loss of accuracy. Note that we could not get PLASQ to work with our data and no software is available to implement CARAT. The preprocessing used by dChip is very similar to CNAT and thus we expect results to be the same. Keep in mind the smoothing step is not being assessed. We observe that the degree of improvement is not equivalent for copy numbers 3 and copy number 1. This is expected because it is easier to detect a 2 times difference (copy number 2 versus 1) than to detect a 1.5 times difference (copy number 3 versus 2).

The most useful application of our results is that we provide improved single-point copy number estimates with reliable uncertainty assessment without the need to re-calibrate for new samples. Note that we can easily control our false positive rate by simply restricting calls to SNPs with posterior probabilities close to 1. Figure 4C demonstrates that we can get usable single-point copy number estimates for a large amount of SNPs. Notice that the worst performance is observed for CN=3. This is likely due to the fact that we used model (1) to extrapolate (as done by CARAT).

6 Discussion

We have presented a mixture model approach that permits us to obtain improved copy number estimate as well as reliable single-point copy number calls. A major advantage of our methodology over the best existing one, e.g. CARAT and PLASQ, is that we explicitly model the conditional joint distribution of the intensities given the copy number values. This permits us to model the strong correlation that sometimes exists between A and B and exploit this information to improve bottom-line results. This advantage is best exemplified by Figure 3C where the **C** = (2, 1), (1, 1), and (1, 2) are usefully separable only if we take this correlation into account. Furthermore, avoiding the linearity assumption made by these other procedures seems to help as well. This is best demonstrated by the fact that we perform worst in cases where we rely on this assumption, i.e. making calls for CN = 3. Finally, because we use training data to fit the mixture models, the procedure is entirely linear. Other procedures, such as CARAT and PLASQ rely on non-linear algorithms that present many practical problems.

We have plans to extend and improve our approach in various ways. First, we plan to implement it for the more recent chips. Second, we believe this approach can be used with Illumina's SNP array and thus plan to try it with data from this platform. Third, we plan to add another level to the model that will permit us to borrow strength across the thousands of SNPs to better estimate the parameters of the conditional probabilities. We plan to use an approach similar to that of CRLMM. Fourth, we plan to look for ways to avoid using the linearity assumption to infer the parameter of conditional distributions when $C > 2$. We plan to use general regression approaches that predict these parameters from the known parameters $C <= 2$. We can train this regression model with *trisomy 21* data ($C = 3$) and design experiments to be able to train for $C > 3$. Fifth, we plan to look for better ways of combining the results form sense and antisense probes. It is desirable to detect and ignore misbehaving strands. Finally, we have observed correlation between parameter estimates coming from proximal locations on the chromosome. This could be due to the fact that various SNPs are on each of the fragments that are amplified. We will explore ways to exploit this finding.

It is possible that the reference set we use has an influence on our results. We plan to study this problem in more detail in the near future. We also plan to substantially increase the size of the reference set to reduce the effect of outlier samples. By combining various publicly available assessment experiments, we plan to develop an comparison protocol for analysis methods. This will help us determine not only which methods work better, but to explore if subsets of the reference set provide better results.

Notice that we did not offer any solutions for the smoothing step as we are more interested in developing techniques for single-point estimates. We expect some of the existing techniques to work well when applied to our estimates of copy number. However, because we explicitly model the conditional probabilities it is possible to develop new methods that impose the across-chromosome correlation through those probabilities instead of the actual copy number estimates.

Acknowledgements

We thank Terry Speed, Giovanni Parmigiani and Ingo Ruczinski for discussion and suggestions. The work of Wenyi Wang was partially funded by Grant No. R01CA105090-01A1. The work of Rafael A Irizarry was partially funded by 1P41HG004059 and P50 HL73994 (Core E); Benilton Car-valho was funded by Coordenao de Aperfeioamento de Pessoal de Nvel Superior (CAPES/Brazil) Aravinda Chakravarti was supported by NIH grants HG02757, MH60007, and the D.W. Reynolds Foundation.

References

Bignell, Graham R.; Huang, Jing; Greshock, Joel; Watt, Stephen; Butler, Adam; West, Sofie; Grigorova, Mira; Jones, Keith W.; Wei, Wen; Stratton, Michael R.; Futreal, P Andrew; Weber, Barbara; Shapero, Michael H.; Wooster, Richard. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 2004 Feb;14(2):287–295. [PubMed: 14762065]URL http://dx.doi.org/10.1101/gr.2012304.

Carvalho, Benilton; Bengtsson, Henrik; Speed, Terence P.; Irizarry, Rafael A. Exploration, normalization, and genotype calls of high density oligonucleotide SNP array data. Biostatistics. 2006; (111)URL http://biostatistics.oxfordjournals.org/cgi/content/abstract/kxl042v1.

Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998 Dec;8(12):1229–1231. [PubMed: 9872978]

Gribble, Susan M.; Kalaitzopoulos, Dimitrios; Burford, Deborah C.; Prigmore, Elena; Selzer, Rebbeca R.; Ng, Bee L.; Matthews, Nick SW.; Porter, Keith M.; Curley, Rebecca; Lindasy, Sarah J.; Baptista, Julia; Richmond, Todd A.; Carter, Nigel P. Ultra-high resolution array painting facilitates breakpoint sequencing. J Med Genet. 2006 Sep;URL http://dx.doi.org/10.1136/jmg.2006.044909.

Huang, Jing; Wei, Wen; Zhang, Jane; Liu, Guoying; Bignell, Graham R.; Stratton, Michael R.; Futreal, Andrew P.; Wooster, Richard; Jones, Keith W.; Shapero, Michael H. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 2004 May;1(4): 287–299. [PubMed: 15588488]

Huang, Jing; Wei, Wen; Chen, Joyce; Zhang, Jane; Liu, Guoying; Xiaojun, Di; Mei, Rui; Ishikawa, Shumpei; Aburatani, Hiroyuki; Jones, Keith W.; Shapero, Michael H. CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 2006;7:83. [PubMed: 16504045]URL http://dx.doi.org/10.1186/1471-2105-7-83.

Huber W, von Heydebreck A, Sueltmann H, Annemarie Poutska, Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;1

Irizarry R, Hobbs B, Collin F, Beaxer-Barclay Y, Antonellis K, Scherf U, Speed T. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–264. [PubMed: 12925520]

Ishikawa, Shumpei; Komura, Daisuke; Tsuji, Shingo; Nishimura, Kunihiro; Yamamoto, Shogo; Panda, Binaya; Huang, Jing; Fukayama, Masashi; Jones, Keith W.; Aburatani, Hiroyuki. Allelic dosage analysis with genotyping microarrays. Biochem Biophys Res Commun 2005 Aug;333(4):1309–1314. [PubMed: 15982637]URL http://dx.doi.org/10.1016/j.bbrc.2005.06.040.

Kennedy, Giulia C.; Matsuzaki, Hajime; Dong, Shoulian; Liu, Wei min; Huang, Jing; Liu, Guoying; Su, Xing; Cao, Manqiu; Chen, Wenwei; Zhang, Jane; Liu, Weiwei; Yang, Geoffrey; Di, Xiaojun; Ryder, Thomas; He, Zhijun; Surti, Urvashi; Phillips, Michael S.; Boyce-Jacino, Michael T.; Fodor, Stephen PA.; Jones, Keith W. Large-scale genotyping of complex DNA. Nat Biotechnol 2003;21:1233–1237. [PubMed: 12960966]

Komura, Daisuke; Nishimura, Kunihiro; Ishikawa, Shumpei; Panda, Binaya; Huang, Jing; Nakamura, Hiroshi; Ihara, Sigeo; Hirose, Michitaka; Jones, Keith W.; Aburatani, Hiroyuki. Noise Reduction from genotyping microarrays using probe level information. In Silico Biol 2006 Feb;6(1–2):0009.

Laframboise, Thomas; Harrington, David; Weir, Barbara A. PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage in Cancer Cells from SNP Array Data. Biostatistics. 2006 Jun;URL http://dx.doi.org/10.1093/biostatistics/kxl012.

Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Science U S A 2001;98:31–36.

Nannya, Yasuhito; Sanada, Masashi; Nakazaki, Kumi; Hosoya, Noriko; Wang, Lili; Hangaishi, Akira; Kurokawa, Mineo; Chiba, Shigeru; Bailey, Dione K.; Kennedy, Giulia C.; Ogawa, Seishi. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 2005 Jul;65(14):6071–6079. [PubMed: 16024607] URL http://dx.doi.org/10.1158/0008-5472.CAN-05-0465.

Peiffer, Daniel A.; Le, Jennie M.; Steemers, Frank J.; Chang, Weihua; Jenniges, Tony; Garcia, Francisco; Haden, Kirt; Li, Jiangzhen; Shaw, Chad A.; Belmont, John; Cheung, Sau Wai; Shen, Richard M.; Barker, David L.; Gunderson, Kevin L. High-resolution genomic profiling of chromosomal

aberrations using Infinium whole-genome genotyping. Genome Res 2006 Sep;16(9):1136–1148. [PubMed: 16899659]URL http://dx.doi.org/10.1101/gr.5402306.

Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 2006 Jan; 22(1):7–12. [PubMed: 16267090]URL http://www.hubmed.org/display.cgi?uids=16267090.

Rocke DM, Durbin B. A model for measurement error for gene expression arrays. J Comput Biol 2001;8 (6):557–569. [PubMed: 11747612]URL http://www.hubmed.org/display.cgi?uids=11747612.

Sharp, Andrew J.; Hansen, Sierra; Selzer, Rebecca R.; Cheng, Ze; Regan, Regina; Hurst, Jane A.; Stewart, Helen; Price, Sue M.; Blair, Edward; Hennekam, Raoul C.; Fitzpatrick, Carrie A.; Segraves, Rick; Richmond, Todd A.; Guiver, Cheryl; Albertson, Donna G.; Pinkel, Daniel; Eis, Peggy S.; Schwartz, Stuart; Knight, Samantha JL.; Eichler, Evan E. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 2006 Sep;38(9):1038–1042. [PubMed: 16906162]URL http://dx.doi.org/10.1038/ng1862.

Wu, Zhijin; Irizarry, RA.; Gentleman, R.; Martinez-Murillo, F.; Spencer, F. A model based back-ground adjustement for oligonucleotide expression arrays. Journal of the America Statistical Association. 2004

Zhao, Xiaojun; Li, Cheng; Guillermo Paez, J.; Chin, Koei; Jeanne, Pasi A.; Chen, Tzu-Hsiu; Girard, Luc; Minna, John; Christiani, David; Leo, Chris; Gray, Joe W.; Sellers, William R.; Meyerson, Matthew. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004 May;64(9):3060–3071. [PubMed: 15126342]

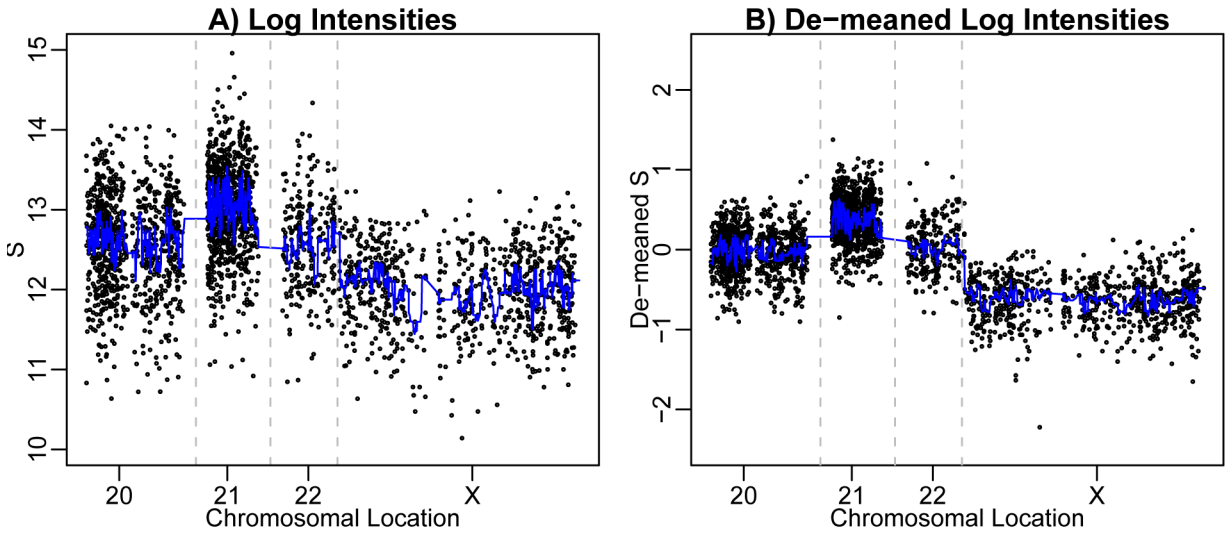

Figure 1.
Log (base 2) intensity, $S_{i,j}$ plotted against chromosomal position of SNP *i*. Chromosomes 20, 21, 22, and X are shown for a male with Down syndrome. Notice that we expect values from chromosome 21 and X to be higher (3 copies) and lower (1 copy) respectively.

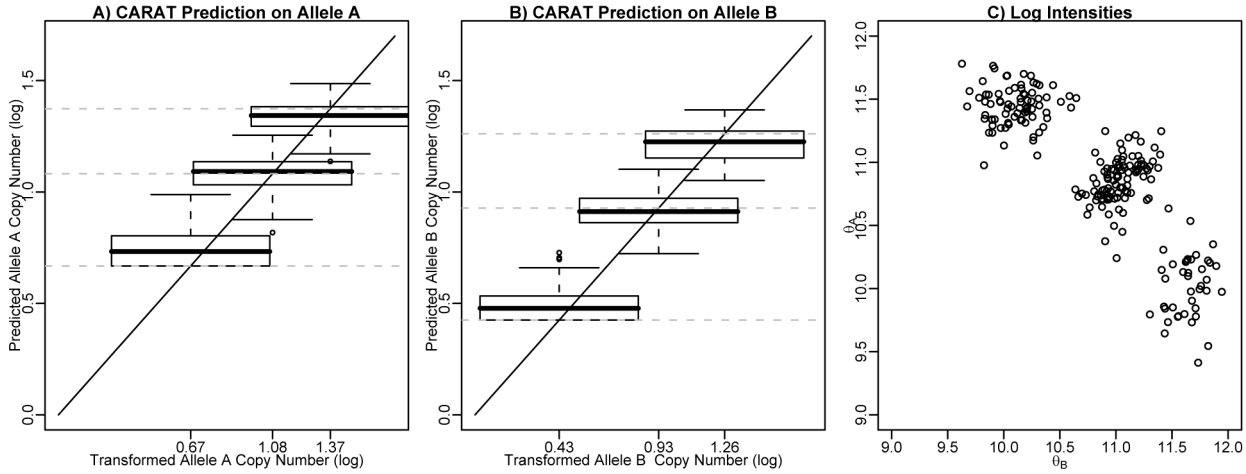

Figure 2.

CARAT regression prediction on a SNP from Chromosome X. The dashed grey lines indicate allele specific copy number = 0,1,2 (from low to high). Figure A) and B) shows the allele specific prediction of copy number (log base 2) as compared to the real copy number (log base 2). We can see that even though the middle of each boxplot lies closely to the intersection of the dashed lines and the diagonal line, the ranges of boxplots overlap. Figure C) shows a scatterplot of preprocessed log intensities for allele A and B.

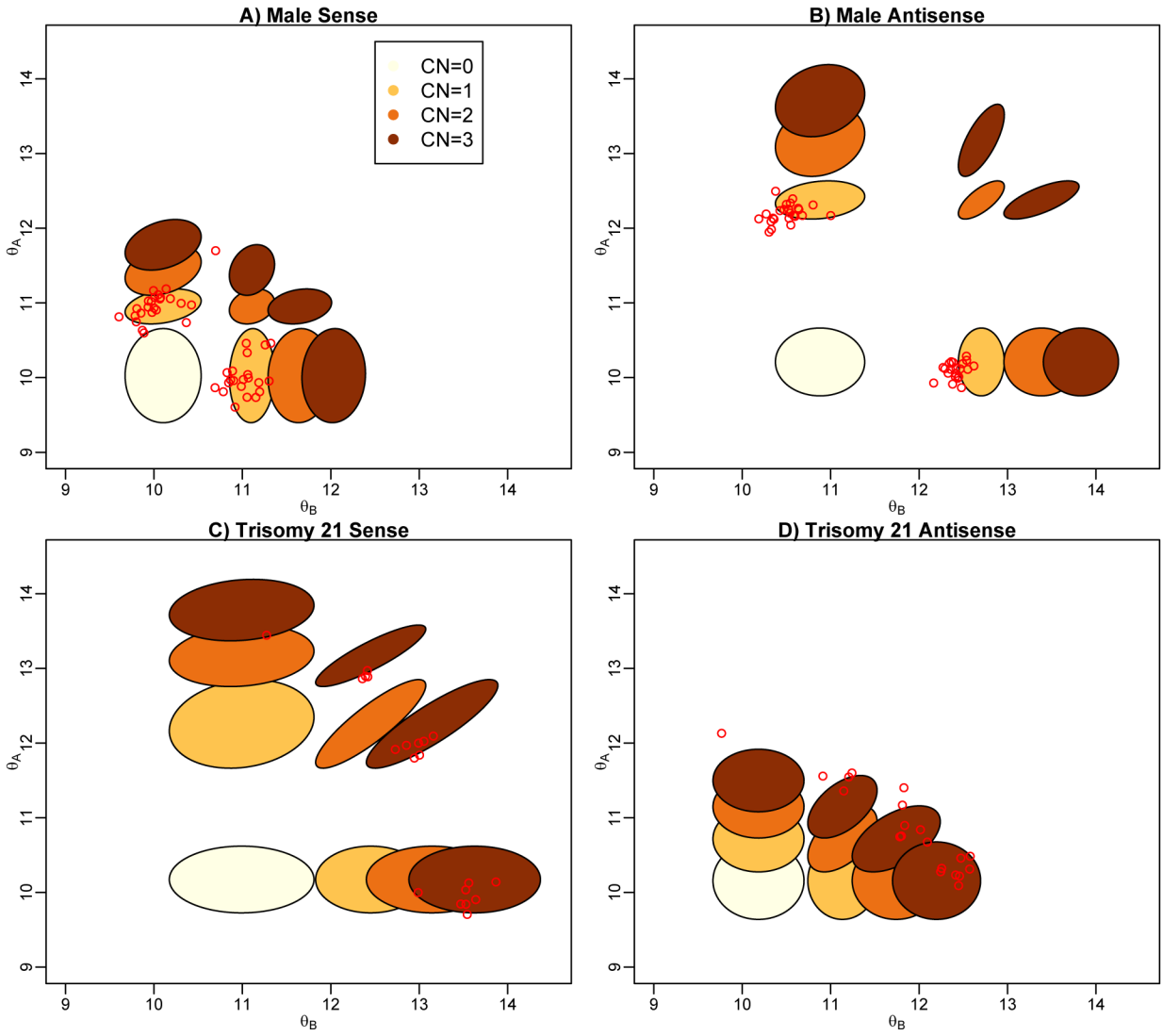

Figure 3.

Conditional joint distributions of $[\theta_{i,j}|\mathbf{C}_{i,j} = \mathbf{c}]$ for a SNP on chromosome X and a SNP on chromosome 21. The X-axis is the log base 2 intensity for allele B. The Y-axis is the log base 2 intensity for allele A. The red dots are from 45 new samples of male normal and 20 Down syndrome patients, respectively. The ellipses shows the 95% critical region around the centers. The brown ellipses represent $C = 3$, $\mathbf{C}_{i,j} = (0, 3), (1, 2), (2, 1), (3, 0)$. The orange ellipses represent $C = 2$, $\mathbf{C}_{i,j} = (0, 2), (1, 1), (2, 0)$. The tan ellipses represent $C = 1$, $\mathbf{C}_{i,j} = (0, 1), (1, 0)$. The yellow ellipses represent $C = 0$, $\mathbf{C}_{i,j} = (0, 0)$.

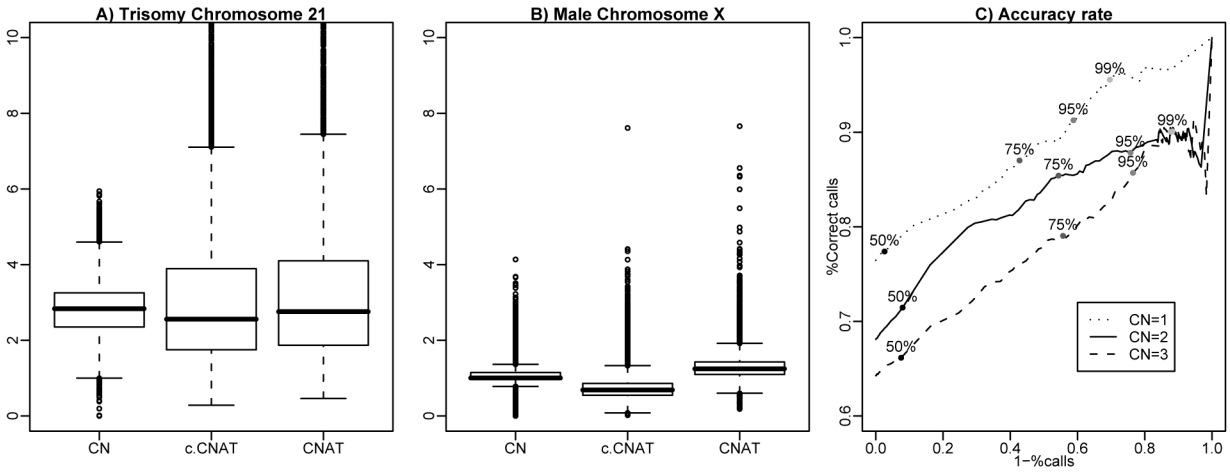

Figure 4.

CN-RLMM results. CN is abbreviation for CN-RLMM, c.CNAT is CRLMM preprocessed probes + CNAT. Figure A) shows the expected copy number given preprocessed log intensities for 817 SNPs on Chromosome 21 of 20 Down syndrome patients (with identified trisomy 21). Figure B) shows the expected copy number given preprocessed log intensities for 807 SNPs on Chromosome X of 45 male trio samples. Figure C) shows the average true positive rate versus 1-average percentage of call rate for 2 Down syndrome patients. The points demonstrate some of the corresponding posterior probability values used as cut-offs.